Birthday Paradox

What's the least number of people in a room such that the probability that some two people share a birthday is greater than 50%?

Assume that there are 365 possible birthdays and that each of these is equally likely.

Let there be k people in the room. Simulate going around and asking everyone what their birthday is…

The probability that each person's birthday is unique is:

$$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \ldots \times \frac{365 - k + 1}{365} < .5$$

For each person, multiply in the probability that their birthday is different than everyone we previously asked…

Goal find min(k) that satisfies that equation.

Write a program =)

```
N = 365
K = 100
pUnique = 1.0
for i in range(N, N-K, -1):
    pUnique = pUnique*i/N
    print(N-i+1, 1-pUnique)
```

It's called the birthday paradox because most people don't think that 23 people is all you need…they think it's much more…

The reason is that although there are only 23 people, we aren't trying to match a specific birthday. Rather, we are looking for any match.

This is a very different question than: Get 23 people in a room. Do any of them have MY birthday?

In this question, we are really asking if any of $\binom{23}{2}$ pairs have the same birthday or not. And so there are a lot more pairs to consider than single people matching a single birthday…

Agenda
----
0) Birthday Paradox
1) Analysis of Craps
2) Discrete Random Variables
3) Binomial Distribution
4) Next time - Continuous Random Variables

Craps

Roll a pair of dice.
If you get 7 or 11, you win
If you get 2, 3, 12, you lose.
Otherwise, let the sum of your roll equal P, which is called the point.

In this case, continue rolling, until you either roll a sum of P or a sum of 7. If you get P first, you win. If you get 7 first, you lose.

Roll sequence: 4, 8, 9, 12, 8, 3, 2, 7 (lose)
Roll sequence: 10, 2, 8, 12, 10 (win)

Question: What is your probability of winning?

For rolling 2 dice here are our probabilities:

2	3	4	5	6	7	8	9	10	11	12
1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

P(Win) = P(7) + P(11) + P(point…getting it again)
= 1/6 + 1/18 + 2*p(4)*p(4 first| 4) + 2*p(5)*p(5 first| 5) + 2*p(6)*p(6 first| 6)
= 1/6 + 1/18 + 2*(3/36)*p(4 first| 4) + 2*(4/36)*p(5 first| 5) + 2*(5/36)*p(6 first| 6)

The *2 for the last three terms are because the probability for 4 is the same as 10 so we're adding 4's probability with 10's probability when we multiply by 2.

We want to know is given some number k and 7 and we keep on rolling, what's the probability we will get k first? Let the probability that we roll k on a single roll be q. (We will be plugging in q = 3/36, 4/36 and 5/36 later…)

Prob rolling 7 = 1/6

P(getting k first) = p(k) +
p(not k or 7)*p(k) +
p(not k or 7)* p(not k or 7)*p(k) +
infinite series with a common ratio of 1 - 1/6 - q
first term is q

Sum = $\frac{q}{1-(1-\frac{1}{6}-q)} = \frac{q}{\frac{1}{6}+q}$

$\frac{3/36}{\frac{1}{6}+\frac{3}{36}} = \frac{3}{9} = \frac{1}{3}$, k = 4, 10

$\frac{4/36}{\frac{1}{6}+\frac{4}{36}} = \frac{4}{10} = \frac{2}{5}$, k = 5, 9

$\frac{5/36}{\frac{1}{6}+\frac{4}{36}} = \frac{5}{11}$, k = 6,8

= 1/6 + 1/18 + 2*(3/36)*(1/3) + 2*(4/36)*(2/5) + 2*(5/36)*(5/11)

$$=\frac{2}{9} + \frac{10+16}{36\times5} + \frac{50}{36\times11} = \frac{4}{18} + \frac{13}{18\times5} + \frac{25}{18\times11} = \frac{220+143+125}{18\times5\times11} = \frac{488}{90\times11} = \frac{244}{495}$$

Non-infinite series solution:

Probability of one event is q, probability of other event is 1/6, keep on repeating until one or the other occurs:

Let X be the answer to the question:

X = p(winning in 1 term) + p(winning on a later turn)

X = q + (1 - 1/6 - q)X
X - (1 - 1/6 - q)X = q
X(1 - (1 - 1/6 - q)) = q
X = $\frac{q}{\frac{1}{6}+q}$

<u>Discrete Random Variables</u>
This is a variable that takes on different values with different probabilities.

For pair of dice rolls we would say

X = 2 with probability 1/36
X = 3 with probability 2/36
…

X = 12 with probability 1/36

Essentially, this is a discrete random variable:

2	3	4	5	6	7	8	9	10	11	12
1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Definitions for Discrete Random Variables:

Expectation (Average Value)

E(X) = $\sum_{x \in X} (p_x) x$

$$\frac{1}{36} \times 2 + \frac{2}{36} \times 3 + \frac{3}{36} \times 4 + \cdots + \frac{1}{36} \times 12$$

$$\frac{2 + 6 + 12 + 20 + 30 + 42 + 40 + 36 + 30 + 22 + 12}{36} = \frac{252}{36} = 7$$

Smaller example:

X = 1, w/prob 1/6
= 2 w/prob 1/3
= 5 w/prob 1/2

E(X) = 1 x 1/6 + 2 x 1/3 + 5 x 1/2 = 20/6 = 10/3

a) Lottery Example
b) Linearity of Expectation
c) Standard Deviation

Nov 11 Lottery
Jackpot = 13.5 million
Match 5 = 6000

Match 4 = 100
Match 3 = 10

N = 53 numbers choose from

P(6) = $\frac{1}{\binom{53}{6}}$, P(5) = $\frac{\binom{6}{5}\binom{46}{1}}{\binom{53}{6}}$, P(4) = $\frac{\binom{6}{4}\binom{46}{2}}{\binom{53}{6}}$, P(3) = P(5) = $\frac{\binom{6}{3}\binom{46}{3}}{\binom{53}{6}}$,

$$\frac{13500000}{\binom{53}{6}} + \frac{6000\binom{6}{5}\binom{46}{1}}{\binom{53}{6}} + \frac{100\binom{6}{4}\binom{46}{2}}{\binom{53}{6}} + \frac{10\binom{6}{3}\binom{46}{3}}{\binom{53}{6}}$$

$$\sim 0.8600464859383521$$

So, if you spend \$1, you would expect to receive 86 cents back. But… of course, the standard deviation is high…vast majority of people lose those 86 cents on their ticket and most of those losses accumulate to very, very few people.

<u>Linearity of Expectation</u>
If X and Y are two discrete random variables, then

E(X) + E(Y) = E(X+Y), (no independence needed for this)

If X and Y are independent E(X)E(Y) = E(XY).

Helpful with dice rolls…treat X as the discrete random variable for die1, Y as the discrete random variable for die2. E(X) = 7/2 (work this out for yourself), E(Y) = 7/2, so that means that the expected value for the sum of two dice rolls is just 7/2 + 7/2 = 7

X = 10,000 p=1/2	Y = 20,000 p = 1/3
15,000 p =1/4	30,000 p = 1/2

18,000 p = 1/4 50,000 p = 1/6

Expected value of both investments together is E(X) + E(Y)

Variance:

Given a discrete random variable X with E(X), its variance is as follows:

$$Var(X) = \sum_{x \in X} p_x(x - E(X))^2$$

Standard Deviation is just the square root of variance.

$$Var(X) = E(X^2) - [E(X)]^2$$

FYI, E(X^2) = $\sum_{x \in X}(p_x)x^2$

$$\sum_{x \in X} p_x(x - E(X))^2 = \sum_{x \in X}(p_x x^2 - 2p_x x E(X) + p_x[E(X)]^2)$$

$$= \sum_{x \in X} p_x x^2 - \sum_{x \in X} 2p_x x E(X) + \sum_{x \in X} p_x[E(X)]^2$$

$$= E(X^2) - 2E(X)\sum_{x \in X}(p_x)x + [E(X)]^2 \sum_{x \in X} p_x$$

$$= E(X^2) - 2E(X)E(X) + [E(X)]^2 \sum_{x \in X} p_x$$

$$= E(X^2) - 2E(X)E(X) + [E(X)]^2$$

$$= E(X^2) - [E(X)]^2$$